

An Empirical Evaluation of GitHub Copilot's Code Suggestions

Year: 2022 | Citations: 323 | Authors: N. Nguyen, Sarah Nadi

Abstract

GitHub and OpenAI recently launched Copilot, an “AI pair programmer” that utilizes the power of Natural Language Processing, Static Analysis, Code Synthesis, and Artificial Intelligence. Given a natural language description of the target functionality, Copilot can generate corresponding code in several programming languages. In this paper, we perform an empirical study to evaluate the correctness and understandability of Copilot’s suggested code. We use 33 LeetCode questions to create queries for Copilot in four different programming languages. We evaluate the correctness of the corresponding 132 Copilot solutions by running LeetCode’s provided tests, and evaluate understandability using SonarQube’s cyclomatic complexity and cognitive complexity metrics. We find that Copilot’s Java suggestions have the highest correctness score (57%) while JavaScript is the lowest (27%). Overall, Copilot’s suggestions have low complexity with no notable differences between the programming languages. We also find some potential Copilot shortcomings, such as generating code that can be further simplified and code that relies on undefined helper methods.